DESIGNING AN ESSAY ANNOTATION RUBRIC AND MODELING ESSAY ORGANIZATION USING STATISTICAL MACHINE LEARNING

by

ALAN C. DAVIS, B.S.

APPROVED BY SUPERVISORY COMMITTEE:

Dr. Vincent Ng, Chair

Dr. Yang Liu

Dr. Haim Schweitzer

To my enlightening and endlessly patient advisor, Dr. Vincent Ng.

I still don't know why he has put up with me after all these years.

DESIGNING AN ESSAY ANNOTATION RUBRIC AND MODELING ESSAY ORGANIZATION USING STATISTICAL MACHINE LEARNING

by

ALAN C. DAVIS, B.S.

THESIS

Presented to the Faculty of

The University of Texas at Dallas

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

THE UNIVERSITY OF TEXAS AT DALLAS

May 2010

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my research advisor, Dr. Vincent Ng, for his encouragement, guidance, support, and patience throughout the time I have been his student. He has constantly challenged my abilities to think creatively by stretching my mind during his difficult exams. From my first course in Computer Science through my senior honors project, and from my first research as a graduate student through the completion of this Master's thesis, Dr. Ng has always been extremely patient, flexible, and friendly. I will never forget the very first time that I entered his office, and I look back fondly on all the times that we talked late into the night.

Next, I want to thank my research partner, Isaac Persing. Without him, none of the ideas described here would have come to fruition. I appreciate all the time he spent discussing research ideas, developing our large code base, designing experiments to test our system, and running those experiments in the final hours before deadlines. He also did a wonderful job helping us to write the paper submission that formed the basis for the second half of this thesis.

Our six annotators, Andrew Hubbs, Karin Khoo, Jayne Koath, Christopher 'Kit' Maier, Andrew Mallon, and Cory Thornton deserve numerous thanks. Without all the countless hours they each spent annotating hundreds of essays, none of the research described in this thesis would have been possible.

I also thank our writing expert, Numair Choudhury, for his time annotating essays during the infancy of this project and for his valuable feedback on our essay annotation rubric.

I extend thanks to Arie Litovsky, an undergraduate student who also helped in designing the essay annotation rubric and who developed the software which our aforementioned annotators used to annotate the essays.

I also thank my research lab-mate, Kazi Saidul Hasan, for framing an early idea I had for labeling paragraphs and sentences as the layered hidden Markov model described in Chapter 3 and for his assistance with essay distribution.

I extend my gratitude to Dr. Yang Liu and Dr. Haim Schweitzer, who formed the remainder of the supervising committee, for their time and flexibility.

I also thank Dr. Vasileios Hatzivassiloglou, because many of the methods and ideas in this thesis were inspired by material I learned from his courses.

We are indebted to the Computer Science Department at the University of Texas at Dallas for compensating the annotators for the their time, without which we would not have the annotated corpus that was built during this research.

Finally, I extend my appreciation to my family and friends, who have been patient and supportive throughout the time I spent doing research and writing this thesis. I know we will look back fondly on these years, thinking about how wonderful it is that I will never have to repeat them.

April 29, 2010

DESIGNING AN ESSAY ANNOTATION RUBRIC

AND MODELING ESSAY ORGANIZATION

USING STATISTICAL MACHINE LEARNING

Publication No. _____

Alan C. Davis, M.S.
The University of Texas at Dallas, 2010

Supervising Professor: Dr. Vincent Ng

In this thesis, we describe the relatively unstudied problem of automatically scoring the organization of student essays. Since progress in modeling organization is hindered in part by the lack of a publicly available annotated corpus, we first develop a detailed essay annotation rubric. Our rubric is used to evaluate ten distinct dimensions of essay quality and label the discourse function of sentences. To stimulate further work on this problem, we annotate a corpus of student essays and make our corpus available to other researchers. We then design a computational model for scoring the organization of student essays. Our model labels discourse structures at both the local and global levels of a test essay and then relates these structures to previously seen paragraphs and essays to evaluate the test essay's organization. Experimental results show that our scoring system significantly outperforms a baseline system.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Automated essay scoring, the task of employing computer technology to evaluate and score written text, is one of the most important educational applications of natural language processing (NLP) (see [1] and [2] for an overview of the state of the art in this task). Recent years have seen a surge of interest in this and other educational applications in the NLP community, as evidenced by the panel discussion on "Emerging Application Areas in Computational Linguistics" at NAACL 2009, as well as increased participation in the series of workshops on "Innovative Use of NLP for Building Educational Applications". Besides its potential commercial value, automated essay scoring brings about a number of relatively less-studied discourse-level problems that involve the computational modeling of different facets of text structure, such as content, coherence, and organization.

A major weakness of many existing scoring engines such as E-rater [3] and the Intelligent Essay Assessor$^{\text{TM}}$[4] is that they adopt a holistic scoring scheme, which summarizes the quality of an essay with a single score and thus provides very limited feedback to the writer. In particular, it is not clear to which dimension of an essay (e.g., coherence, relevance, development) a score should be attributed. Recent work addresses this problem by scoring a particular dimension of essay quality such as coherence [5], technical errors, and relevance to prompt [6].

To our knowledge, there is an essay scoring dimension for which no computational model has yet been developed — *organization*. Organization refers to the structure of an essay. A high score on organization means that writers introduce a

topic, state their position on that topic, support their position with main ideas and evidence, and then conclude, often by restating their position [7]. A well-organized essay is structured in a way that logically develops an argument. Note that organization is a different facet of text structure than coherence. While coherence refers to transitions between the *content* and *ideas* present in subsequent sentences, organization refers to transitions between the *functions* of sentences and paragraphs and how these are arranged together to structure an argument.

While many models of text coherence have been developed in recent years (e.g., [8], [9], [10], [11]), the same is not true for text organization. One reason is the availability of training and test data for coherence modeling. Coherence models are typically evaluated on the sentence ordering task, and hence training and test data can be generated simply by scrambling the order of the sentences in a text. On the other hand, it is not particularly easy to find poorly organized texts for training and evaluating organization models. We believe that student essays are an ideal source of well- and poorly-organized texts. Thus, we have found the International Corpus of Learner English (ICLE) [12] to be an untapped source of student essays.

The first goal of this thesis is to build an annotated corpus of student essays with which we can train computational models that evaluate the quality of essays. To do this, we first develop a detailed essay annotation rubric that is used to examine ten distinct dimensions of essay quality and label the discourse function of sentences. We then annotate a corpus of student essays selected from the ICLE. Unlike previous work where the corpus is annotated with a binary decision (i.e., *good* or *bad*) for a given scoring dimension (e.g., [6]), we annotate each essay in our corpus with one of seven scores. This not only provides a finer-grained distinction of essay quality (which can be important in practice), but also makes the prediction tasks more challenging.

The second goal of this thesis is to develop a computational model for scoring the organization of student essays. Our model labels discourse structures at both the global (i.e., paragraph) and local (i.e., sentence) levels of a test essay and then relates these structures to previously seen essays and paragraphs to evaluate the test essay's organization. Experimental results show that our scoring system significantly outperforms a baseline system.

In sum, the contributions presented to the research community in this thesis are two-fold. First, we develop a detailed essay annotation rubric with which we annotate our essay corpus, and we make this data set available to the public. Since progress in modeling some dimensions of essay quality (organization, etc.) is hindered in part by the lack of a publicly available annotated corpus, we believe that our data set will be a valuable resource to the NLP community. Second, we address a less-studied discourse-level task — predicting the organization score of student essays — by developing a computational model of organization. With this new annotated corpus and organization model, we hope to stimulate further interest in this problem.

CHAPTER 2
CORPUS ANNOTATION

Our first goal in this research was to build an annotated corpus of student essays. To do this, we first chose a reputable corpus which contains student essays, then we developed an essay annotation rubric which was used to annotate the selected essays. This chapter describes our reasoning as we went through these processes.

2.1 Corpus Selection

We chose to use as our corpus the 4.5 million word International Corpus of Learner English (ICLE) [12], which consists of more than 6000 essays written by university undergraduates from 16 countries and 16 native languages who are learners of English as a Foreign Language. All essays are unabridged with an average length of 617 words.

We selected a subset consisting of 1025 essays from the ICLE to annotate and use for training and testing of automated essay scoring models. While *narrative* writing asks students to compose descriptive stories, *argumentative* (also known as *persuasive*) writing requires students to state their opinion on a given topic and to validate that opinion with a convincing argument. 91% of the essays in the ICLE are argumentative, making this corpus ideal for discourse-oriented exploration. Thus, we selected only argumentative essays because they contain the rhetorical discourse structures we are interesting in modeling.

To ensure representation across the native languages of the authors, we selected essays written in response to topics which are well-represented in multiple languages. This avoids many issues that may arise when certain vocabulary is used in response

to a topic for which essays written by authors from only a few languages are available. Table 2.1 shows the ten topics we selected for annotation, and Table 2.2 shows the distribution of prompts and essays across the fifteen native languages represented in this set of essays.

Table 2.1. Writing topics selected for annotation

Topic	Languages	Essays
Some people say that in our modern world, dominated by science, technology and industrialisation, there is no longer a place for dreaming and imagination. What is your opinion?	12	298
Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.	13	148
The prison system is outdated. No civilized society should punish its criminals; it should rehabilitate them.	11	98
In his novel *Animal Farm*, George Orwell wrote "All men are equal but some are more equal than others." How true is this today?	10	82
In the words of the old song: "Money is the root of all evil."	10	75
All armies should consist entirely of professional soldiers; there is no value in a system of military service.	9	68
Feminists have done more harm to the cause of women than good.	11	66
Television is the opium of the masses in modern society.	2	58
Most university degrees are theoretical and do not prepare us for real life. Do you agree or disagree?	1	48
Crime does not pay.	10	39

2.2 Essay Annotation Rubric

We developed an essay annotation rubric by studying rubrics used to score essays written for standardized tests (e.g., SAT and GRE) as well as annotation protocols created for the purpose of developing computational models of language (e.g., [13]). We analyzed the strengths and weaknesses of each of these rubrics and protocols so that we could design our own rubric that addresses our concerns with the others.

Table 2.2. Distribution of essays and topics across authors' native languages

Language	Prompts	Essays
Bulgarian	3	145
Norwegian	7	102
French	8	95
Dutch	8	83
Russian	8	79
Spanish	8	73
Mandarin	4	70
Turkish	2	66
Tswana	5	55
Finnish	8	52
Swedish	5	50
Italian	8	43
Czech	7	38
Polish	6	15
Cantonese	2	14

Throughout the process of developing our rubric, we consulted a writing expert with years of experience in grading and proofreading essays, who evaluated essays using early versions of our rubric and provided us with feedback on how to improve it.

2.2.1 Dimensions for Scoring Essay Quality

Some rubrics aim to summarize the overall quality of an essay with one numerical score. However, we found that this holistic approach lacks the ability to provide feedback about what specific qualities or issues cause an essay to receive a high or low score. For this reason, we identified ten distinct dimensions of essay quality — Adherence to Prompt, Clarity of Thesis, Strength of Argument, Development, Organization, Coherence, Cohesion, Sentence Structure, Vocabulary, and Technical Quality — which are scored independently. We defined each dimension to evaluate a specific aspect of essay quality to avoid overlap between dimensions. Our writing expert helped us to make clearer distinctions between dimensions which seemed too

similar. We experimented with scoring some of the dimensions at the sentence level (e.g., Technical Quality) and others at the paragraph level (e.g., Development), but ultimately realized that all dimensions should be scored at the document level.

Human annotators scored the essays we selected using written guidelines in our essay scoring rubric. Annotators evaluated each dimension of essay quality using a numerical score from 1 to 4 at half-point increments. We used this scoring system so that it is easy to interpret the numerical values into four categories of quality. Scores of 1, 2, 3, and 4 indicate that an essay is terrible, poor, good, and excellent with respect to a particular dimension of quality. By using an even number of score values, we allow the possibility of developing models or running experiments in which scores are collapsed into binary values (i.e., *good* or *bad*). The ability to do this may be of interest if one wants to treat essay scoring as a binary classification task.

However, we later decided to allow half-point scores because we found that occasionally two annotators who think an essay's score should fall somewhere between two integers would each decide to score it with different values. By allowing half-point scores, we saw higher agreement between annotators in cases like this. For example, two annotators may both think an essay is fairly well organized, but one decides to assign it a score of 3 while the other gives it a score of 4. If half-point scores are allowed, in this case both annotators may decide to assign the essay a score of 3.5 on the Organization dimension.

The ten dimensions of essay quality are described in our essay scoring rubric and were discussed in detail during meetings to discuss sample essays. These sample essays allowed annotators to see examples of both good and bad writing with respect to different dimensions of quality, and the meetings to discuss the samples effectively served to calibrate annotators. Annotators maintained their calibration by following

the descriptions given in the rubric closely. We will now proceed to describe the ten dimensions of essay quality and the meaning of each integer score for that dimension.

Adherence to Prompt refers to the relevance of the essay's content. A high score on this dimension means that the essay fully addresses any questions posed by the prompt and the author consistently stays on topic, rarely discussing material not relevant to the argument. Table 2.3 shows the guidelines describing each score for the Adherence to Prompt dimension.

Table 2.3. Essay scoring rubric for the Adherence to Prompt dimension

Score	Description of Adherence to Prompt
4	essay fully addresses the prompt and consistently stays on topic
3	essay mostly addresses the prompt or occasionally wanders off topic
2	essay does not fully address the prompt or consistently wanders off topic
1	essay does not address the prompt at all or is completely off topic

Clarity of Thesis refers to how well the thesis statement explains the author's position. A high score on this dimension means that the essay contains an obvious thesis statement which requires little or no clarification of what the author's stance is on the topic. Table 2.4 shows the guidelines describing each score for the Clarity of Thesis dimension.

Table 2.4. Essay scoring rubric for the Clarity of Thesis dimension

Score	Description of Clarity of Thesis
4	essay presents a very clear thesis and requires little or no clarification
3	essay presents a moderately clear thesis but could benefit from some clarification
2	essay presents an unclear thesis and would greatly benefit from further clarification
1	essay presents no thesis of any kind and it is difficult to see what the thesis could be

Strength of Argument refers to how convincing the author's argument is. A high score on this dimension means that the essay makes a strong argument for its thesis and would convince most readers of the author's position. Table 2.5 shows the guidelines describing each score for the Strength of Argument dimension.

Table 2.5. Essay scoring rubric for the Strength of Argument dimension

Score	Description of Strength of Argument
4	essay makes a strong argument for its thesis and would convince most readers
3	essay makes a decent argument for its thesis and could convince some readers
2	essay makes a weak argument for its thesis or sometimes even argues against it
1	essay does not make an argument or it is often unclear what the argument is

Development refers to the depth to which an essay develops its main ideas. A high score on this dimension means that authors state and support their position fully and develop their argument by providing adequate elaboration, and examples. Table 2.6 shows the guidelines describing each score for the Development dimension.

Table 2.6. Essay scoring rubric for the Development dimension

Score	Description of Development
4	essay fully develops its main ideas with adequate elaboration and examples
3	essay develops most of its ideas but could benefit from further elaboration and examples
2	essay does not fully develop its ideas and would greatly benefit from further elaboration
1	essay presents numerous undeveloped ideas with almost no elaboration or examples

Organization refers to the structure of an essay. A high score on Organization means that writers introduce a topic, state their position on that topic, support their position with different main ideas, and conclude, often by restating their position [7]. A well-organized essay is structured in a way that logically develops an argument. Table 2.7 shows the guidelines describing each score for the Organization dimension.

Table 2.7. Essay scoring rubric for the Organization dimension

Score	Description of Organization
4	essay is very well structured and logically develops an organized argument
3	essay is fairly well structured but could somewhat benefit from reorganization
2	essay is poorly structured and would greatly benefit from reorganization
1	essay is completely unstructured and requires major reorganization

Note that Organization is a different facet of text structure than Coherence. While Organization refers to transitions between the *function* of subsequent sentences, Coherence refers to transitions between the *content* of sentences. A high score on this dimension means that the essay is generally understandable because it transitions between ideas in a sensible manner without confusing the reader. Table 2.8 shows the guidelines describing each score for the Coherence dimension.

Table 2.8. Essay scoring rubric for the Coherence dimension

Score	Description of Coherence
4	essay contains sensible transitions between ideas and is very understandable
3	essay contains a few slightly confusing transitions between ideas but is still understandable
2	essay contains multiple confusion transitions because it switches between ideas roughly
1	essay contains few or no transitions and is a highly fragmented collection of separate ideas

Note that Coherence concerns different types of transitions than Cohesion. While Coherence refers to transitions between *concepts and ideas*, Cohesion refers to transition *words and phrases* between segments of text. A high score on this dimension means that the essay contains appropriate transition words and phrases between paragraphs, sentences, and phrases, which link statements and ideas to show their connections and aid understanding. Table 2.9 shows the guidelines describing each score for the Cohesion dimension.

Table 2.9. Essay scoring rubric for the Cohesion dimension

Score	Description of Cohesion
4	essay contains appropriate transition words and phrases between paragraphs, sentences, and phrases, linking statements and ideas to show their connections and aid understanding
3	essay contains some transition words or phrases but could somewhat benefit from their use
2	essay contains few transition words or phrases and would greatly benefit from their use
1	essay contains almost no transitions and requires their use to help understand connections

Sentence Structure refers to the complexity and variety of how sentences are formed in an essay. A high score on this dimension means that the essay contains varied sentence structures of appropriate complexity, which make the flow of writing interesting to the reader. Table 2.10 shows the guidelines describing each score for the Sentence Structure dimension.

Table 2.10. Essay scoring rubric for the Sentence Structure dimension

Score	Description of Sentence Structure
4	essay contains numerous varied sentence structures of appropriate complexity
3	essay contains somewhat varied sentence structures of moderate complexity
2	essay contains limited sentence structures of rather low complexity
1	essay excessively and inappropriately repeats the same simple sentence structures

Vocabulary refers to how advanced the words used in an essay are and whether the author has appropriate diction. A high score on this dimension means that the essay shows appropriate word choice and contains advanced vocabulary. Table 2.11 shows the guidelines describing each score for the Vocabulary dimension.

Technical Quality refers to both the number and the severity of grammatical, mechanical, spelling, and punctuation errors. A high score on this dimension means

Table 2.11. Essay scoring rubric for the Vocabulary dimension

Score	Description of Vocabulary
4	essay shows appropriate word choice and contains advanced vocabulary
3	essay shows appropriate word choice and contains intermediate vocabulary
2	essay shows limited word choice and contains only beginning vocabulary
1	essay excessively and inappropriately repeats the same words and/or phrases

that the essay contains very few technical errors and that these errors do not affect its overall readability. Table 2.12 shows the guidelines describing each score for the Technical Quality dimension.

Table 2.12. Essay scoring rubric for the Technical Quality dimension

Score	Description of Technical Quality
4	essay contains very few technical errors that do not affect its overall readability
3	essay contains some technical errors that make it only somewhat difficult to read
2	essay contains many technical errors that make it significantly difficult to read
1	essay contains numerous technical errors that make it extremely difficult to read

In addition to these ten specific dimensions of quality, we chose to also evaluate the overall essay quality as a way to take everything into account and summarize the scores given to individual dimensions. The Overall score is not simply an average of the other dimensions' scores, since each may have a different level of importance. While each score on the individual dimensions should affect the overall score to some extent, none should have too strong an influence on the Overall score. For example, a coherent and well-developed essay should not be overly penalized for poor grammar. Scores in the 3 to 4 range indicate good to excellent essays, while scores in the 1 to 2 range indicate terrible to poor essays. Half-points are allowed for this score as well.

This description may sound too vague for annotators to evaluate properly, but this is part of the purpose of having an Overall score — we want the annotators to

be able to also give a holistic score for the whole essay based on their intuitive "feel" for how good or bad it is. The weight for how much each dimension's score should affect the Overall score is not explicitly given, as there may be aspects of the essay that are not completely captured in those dimensions or that may not be able to be isolated from their context. This is where annotators can score those elusive aspects.

Furthermore, we may be interested in modeling the relationships that exist between the Overall score and the individual dimensions. This task by itself is an interesting machine learning problem that would be important for building a complete automated essay scoring system.

2.2.2 Sentence Function Labels

The annotation protocol used by Burstein et al. for identifying discourse structure in student essays [13] has human judges label spans of text with discourse categories. The labels are used to segment the essay text into distinct rhetorical structures [14]. We realized that it would be important to identify these types of discourse structures when developing computational models of organization, coherence, development, and other dimensions of quality that concern the structure of an essay.

Thus, in addition to scoring the ten dimensions of quality, annotators labeled each sentence with a discourse function that indicates its logical role in the argument. Our label schema is based on work in discourse structure by Burstein, et al. [13], but features six additional labels which we added to classify sentences whose function is distinct from the other categories. Once again, we consulted our writing expert to help us make clear distinctions between sentence labels which seemed similar and to propose new sentence labels when necessary. We chose intuitive names that are similar to the vocabulary commonly used by writing instructors and English teachers.

Table 2.13 contains descriptions of the twelve sentence function labels used in our essay annotation rubric.

Table 2.13. Descriptions of sentence function labels

Label	Name	Sentence Function
P	Prompt	restates prompt given to author and contains no new material or opinions
B	Background	introduces peripheral material that provides context for other material
T	Transition	shifts focus to new topics but contains no meaningful information
H	Thesis	states author's position on the topic for which he/she is arguing
M	Main Idea	asserts reasons and foundational arguments that support thesis
E	Elaboration	further explains reasons and ideas but contains no evidence or examples
S	Support	provides evidence and examples to support claims made in other statements
C	Conclusion	summarizes and concludes the entire argument or a main ideas
R	Rebuttal	considers counter-arguments that contrast with thesis or main ideas
O	Solution	puts to rest questions and problems brought up by counter-arguments
U	Suggestion	proposes solutions problems brought up by argument
I	Irrelevant	does not meaningfully contribute to argument in a substantive way

2.3 Annotator Selection and Training

We selected our six annotators from more than 30 applicants through a competitive selection process. First, applicants were familiarized with the essay scoring rubric at an orientation meeting and given a sample essay to annotate. Those who maintained consistency with the expected scores and sentence labels were invited back to the next round to discuss their annotations with other applicants. At each meeting, we discussed the sample essay in detail until applicants reached consensus on the best annotations. After four rounds of sample essay scoring, discussion, question answering, and applicant selection, the remaining annotators had been calibrated to each other. The six annotators who were most consistent with the expected scores and sentence

labels and who could justify their annotations with strong arguments were then given approximately 200 essays to annotate. Among these, 45 essays were distributed to all annotators to evaluate inter-annotator agreement and ensure consistency in scoring. The remaining 980 essays were each given to an annotator so that approximately 160 of the essays each annotator received were distinct from those given to the others. Thus, about 22% of the essays distributed to each annotator were overlapping and the other 78% were unique.

It is essential that the annotators have high agreement because this suggests that humans can agree on how to evaluate these ten dimensions of essay quality. We calculate inter-annotator agreement by comparing the scores of all pairs of graders and averaging over all shared essays. Because our scoring system allows for such fine-grained evaluation, we should not reasonably expect two humans to agree exactly. Instead, we should consider two humans' scores to agree when they differ by 0.5 or less. With this in mind, we find that annotators' scores agree 67.3% of the time and are within 1 point of each other 89.5% of the time. Although we have six humans scoring ten different dimensions of essay quality, these statistics suggest that we have high enough agreement that we can reasonably expect computational models to capture the annotators' concept of what each dimension represents.

CHAPTER 3
MODELING ESSAY ORGANIZATION

The second goal of this research was to develop a computational model for scoring the organization of student essays. As mentioned before, organization refers to the structure of an essay. A high score on organization means that writers introduce a topic, state their position on that topic, support their position with main ideas and evidence, and then conclude, often by restating their position. A well-organized essay is structured in a way that logically develops an argument.

3.1 Paragraph Type Labels

Thus, an essay's organization score depends heavily on the sequence of paragraphs that compose it. To this end, we define four labels — Introduction, Body, Rebuttal, and Conclusion — which can be used to categorize each of an essay's paragraphs. Table 3.1 gives descriptions of the four paragraph type labels used in our organization model.

Table 3.1. Descriptions of paragraph type labels

Label	Name	Paragraph Type
I	Introduction	introduces essay topic and states author's position and main ideas
B	Body	provides reasons, evidence, and examples to support main ideas
R	Rebuttal	considers counter-arguments to thesis or main ideas
C	Conclusion	summarizes and concludes arguments made in body paragraphs

To illustrate why these paragraph labels are useful for assigning organization scores, consider the two essays described by the following sequences of paragraphs. Essay one begins with a paragraph Introducing the topic, followed by several Body

paragraphs, and ends with a Conclusion. Essay two contains the same number of paragraphs as essay one, but begins with a Conclusion paragraph followed by several Rebuttals of the Conclusion and ends with an Introduction of the essay's topic. From this information, it appears that essay one's paragraphs follow a reasonable pattern and it should therefore be assigned a high organization score. Essay two follows a much less reasonable pattern, so it should receive a lower score.

3.2 Predicting Paragraph and Sentence Labels

The question that now arises is, how can we assign these type labels to an essay's paragraphs? Just as a sequence of paragraphs can be used to estimate an essay's organization score, a sequence of sentences in a paragraph can be used to assign the paragraph a label. As described in Section 2.2.2, each of a paragraph's sentences can be assigned one of 12 function labels. For example, a paragraph containing Thesis or Prompt sentences is likely to be an Introduction or Conclusion paragraph, whereas one containing Rebuttal or Solution sentences is more likely to be a Body or Rebuttal paragraph. Therefore, before we can accurately determine whether a given paragraph is an example of an Introduction, Body, Conclusion, or Rebuttal, we must first discover the functions of its component sentences. We have identified three categories of features which may be useful in determining a sentence's function.

Unigram word features have been frequently and successfully applied in many text categorization tasks. To motivate our use of Word features for sentence function labeling, consider the words "firstly", "example", and "thus". While "firstly" most frequently occurs in Main Idea sentences, "example" and "thus" usually occur in Support and Conclusion sentences, respectively.

Another variety of textual feature we use to label sentences is the types of transitions that appear in them. We obtained fourteen lists of transitional words and phrases[1], each associated with a different type of transition (e.g., Summarizing transitions, which include the phrases "all things considered" and "in conclusion"). Conclusion sentences are more likely to contain Summarizing phrases such as "in conclusion", but Support sentences more often contain Illustration phrases such as "for example". We therefore defined 15 binary features, 14 indicating the categories of transitions a sentence contains and one indicating whether or not the sentence contains any transitional phrases.

Finally, each sentence in an essay can also be associated with a set of exactly six positional features like those described by Burstein, et al. [13]. The positional features we used include the position of the sentence with respect to (1) the beginning of the essay, (2) the end of the essay, (3) the beginning of the paragraph it appears in, and (4) the end of the paragraph it appears in, as well as its paragraph's position with respect to (5) the beginning of the essay and (6) the end of the essay. So, for example, the fifth sentence in a 20-sentence essay will have the features `SentenceBeginEssay=5` and `SentenceEndEssay=16`. If this sentence is the second sentence in a five-sentence paragraph, it will have the features `SentenceBeginParagraph=2` and `SentenceEndParagraph=4`. Finally, if the paragraph it appears in is the fourth paragraph of a six-paragraph essay, it will have the features `ParagraphBegin=4` and `ParagraphEnd=3`.

To give some examples of why these features might be useful, we note that the first sentence in an essay is frequently a Prompt restatement, and a sentence appearing in the last paragraph of an essay is often a Conclusion statement. Note

[1]Obtained from Study Guides and Strategies, `http://www.studygs.net/wrtstr6.htm`

that these are not six integer-valued features, but instead a much larger number of binary features. We anticipate that integer valued positional features could confuse a learner because, for example, we cannot necessarily expect the second sentence in a paragraph to be more similar to the first sentence than it is to the last.

3.3 Modeling Essay Structure with a Layered Hidden Markov Model

Thus far, we have described features that can be used for predicting paragraph labels as well as features that can be used to predict sentence labels, but we have not described how we use them. One traditional supervised learning approach would be to train one learner to predict sentence function labels, then train another learner to use the output of the first to predict paragraph labels. This approach suffers from the problem that each learner we introduce in the chain of processes ultimately leading to the organization score estimator may introduce additional errors into the training data the final learner uses. Therefore, we prefer to predict the paragraph labels in one step. For this reason, we use a layered hidden Markov model (HMM) [15] to label paragraphs.

Our layered HMM consists of two layers. The paragraph layer applies one of the four paragraph labels to each of an essay's paragraphs. That is, it has paragraph labels as hidden states and sentence function label sequences as its observations. The sentence layer consists of four separate traditional HMMs, one for each paragraph label. The sentence layer HMMs have sentence functions as their hidden states and a set of Word, Position, and Transition features as their observations. The role of the sentence layer is to provide the paragraph layer HMM with the four most probable sentence function sequences, one for each paragraph label.

Because the paragraphs in our dataset do not have annotator-provided labels, the first step in training the two-layer HMM is to assign an initial label to each training set paragraph. To do this, we constructed a set of heuristic rules for guessing a paragraph's label given the function labels of its sentences. For any paragraph, we first assume each label is equally likely. We then examine each of the paragraph's sentences, applying all of the rules that are appropriate for that sentence. To give some examples, one rule increases our confidence that the paragraph's label is Body if the function of the sentence we are examining is Support. Another rule increases our confidence that the paragraph's label is Introduction or Conclusion if the sentence's function is Thesis or Prompt.

These paragraph-labeling heuristics serve to partition the paragraphs in the training set into four groups, one for each paragraph label. We then calculate the parameters of the Sentence layer HMMs using add-delta-smoothed maximum likelihood estimates. While it is obvious how to make a maximum likelihood estimate of $P(s_i|s_{i-1})$, the probability of sentence i's function given that of sentence $i-1$, calculating $P(o_i|s_i)$, the probability of the set of observations o_i given sentence i's function, is more difficult. The problem is that HMMs do not typically use many features, but in our implementation, o_i consists of a set of 215 observations corresponding to 215 presence-or-absence features (100 Word features, 100 Position features, and 15 Transition features), all selected using information gain [16]. To calculate these probabilities, we assume that each feature is independent of every other feature, and that each feature is independent of its paragraph's label given its sentence's label. Hence, we calculate these probabilities using

$$P(o_i|s_i) = \prod_{j=1}^{215} P(o_{ij}|s_i)$$

Now that the sentence layer HMMs have been trained, we can estimate the parameters of the paragraph layer HMM. Because we have applied provisional labels to each paragraph, $P(p_i|p_{i-1})$ is easy to estimate, but $P(S_k^{p_k}|p_k)$, the probability of sentence label sequence S_k given paragraph k's label is more difficult. To do this, we note that the sentence layer HMMs can estimate the probability of the most likely sentence label sequence, given that it occurs in a paragraph with label p and m sentences, all together having observations O, as

$$\max_S P(S|O,p) \propto \max_S \prod_{i=1}^{m} P(o_i|s_i)P(s_i|s_{i-1},p)$$

With all the two-layer HMM's parameters set, we can finally re-estimate the training paragraphs' labels using the formula

$$\operatorname*{argmax}_P P(P|S) \propto \operatorname*{argmax}_P \prod_{i=1}^{n} P(S_i^{p_i}|p_i)P(p_i|p_{i-1})$$

using the Viterbi algorithm [17]. Now, with these newly re-estimated paragraph labels, we iterate the process described above only five times to avoid overfitting. Note that during training, we keep sentence function labels fixed, only using the sentence layer HMMs for estimating $P(S|O,p)$. To apply paragraph labels to test essays, we need to use the Viterbi algorithm in both the sentence and paragraph layer HMMs since we do not know the sentence function labels of a test essay.

3.4 Finding the Nearest Neighbors of an Essay and Its Paragraphs

Having applied labels to each paragraph in an essay, how can we use these labels to predict the essay's score? Recalling that the importance of each label stems not from the label itself, but from the sequence of labels it appears in, we first consider methods involving sequence alignment. Using the Needleman-Wunsch alignment algorithm [18], we can generate a similarity score for any pair of essays by aligning their paragraph sequences.

This gives us what seems like a simple method for assigning organization scores to test essays: Simply examine a test essay's k nearest neighbors among the training set with respect to their similarity scores and use some method to aggregate the nearest neighbors' organization scores into one number, the predicted test essay score.

However, there are some problems with this method. First, it is not clear what kind of scoring function we should use to perform the alignments. In sequence alignment, scoring function $S(i, j)$ tells us how likely it is that symbol i (in our case, a paragraph label) will be substituted with another symbol j. While we expect that in an alignment between high-scoring essays, an Introduction paragraph is most likely to be aligned with another Introduction paragraph, how much worse should the alignment score be if an Introduction paragraph needs to be aligned with a Rebuttal paragraph or replaced with an indel[2]? We solve this problem by heuristically setting the scoring function such that $S(i, j) = 1$ when $i = j$, $S(i, j) = -1$ when $i \neq j$, and also $S(i, -) = S(-, i) = -1$, where '$-$' is an indel.

Next, it is not clear that the k nearest neighbors of an essay will always be similar to it with regards to organization score. While we do expect the alignment scores between good essays with reasonable paragraph sequences to be high, poorly organized essays by their nature have more random paragraph sequences. Hence, we have no intuition about the k nearest neighbors of a poor essay, as it may have as high an alignment score with another poorly organized essay as with a good essay.

Comparing the sequence of paragraph labels in an essay is only considering the ordering of paragraph types at a high level, so we should also examine the order of sentence functions within each paragraph. The intuition is that at least some

[2]In pairwise sequence alignment, an indel indicates that in order to transform one sequence to match another, we must either insert a symbol into one sequence or delete a symbol from the other sequence, hence the term 'indel'.

portion of an essay's organization score can be attributed not to the essay itself, but to the organization of the sentence sequences of its component paragraphs. To address this concern, we apply the same k nearest neighbor method to align the sentence function sequences of an essay's paragraphs with sentence function sequences of paragraphs in the training set. By associating each training set paragraph with its essay's organization score, this gives n sets of k paragraph organization scores to associate with each essay having n paragraphs.

Finally, it is not clear how best to aggregate all of these essay and paragraph nearest neighbor scores into one value for each essay. The most natural ways are to calculate the mean, median, or mode of the k neighbors' scores. A reasonable argument could be made for using each of these functions, and there is no clear winner from an intuitive point of view. Another approach is to consider the alignment score between the paragraph in the test essay and each of its k nearest neighbors. Since the alignment score is high for a pair of similar label sequences, it is natural to find the weighted mean of the neighbors' scores, using the alignment score as the weight. We therefore define the alignment-score-weighted-mean as

$$\frac{\sum_{j=1}^{k} align(P_i, P_j) \times org.essay(P_j)}{\sum_{j=1}^{k} align(P_i, P_j)}$$

where $align(P_i, P_j)$ is the sequence alignment score of paragraphs P_i and P_j, and $org.essay(P_j)$ is the actual organization score of the essay containing paragraph P_j.

Once we have aggregated the scores of the k neighbors for each of the n paragraphs, we want to combine these scores in some way to predict the organization score of the test essay. We could calculate the mean, median, or mode of the n essays' scores, but we similarly cannot intuitively decide which of these is the best way to predict the essay's score. Another reasonable idea based on the observation that longer paragraphs have more content than shorter paragraphs is to calculate the

weighted mean of the n paragraphs' scores, using the length of each paragraph as its corresponding weight. We therefore define the length-weighted-mean as

$$\frac{\sum_{i=1}^{n} length(P_i) \times org.paragraph(P_i)}{\sum_{i=1}^{n} length(P_i)}$$

where $org.paragraph(P_i)$ is the estimated organization score of paragraph P_i in a test essay with n sentences.

3.5 Using a Regressor to Predict an Essay's Organization Score

In the previous subsection, we mentioned that there are several different ways for predicting an essay's organization score given its k nearest neighbors. Two of the most important objections we raised to using any of these methods directly were our lack of confidence that low-scoring essays would have mostly low-scoring neighboring essays and that it is not intuitively clear which of the methods proposed, if any, should be used. Is it possible to construct an automatic scoring system that can answer these objections by (1) accounting for errors resulting from the nearest neighbor algorithm's inability to align poorly organized essays with similar poorly organized essays, and (2) making use of all the scoring methods described?

Our solution to these problems is to use the organization scores obtained by all the proposed methods as features for a support vector machine learner. We use the SVM^{light} [19] implementation of regression support vector machines [20] to train a learner using these features because SVMs have been frequently and successfully applied in natural language processing problems. We believe that treating the organization scores obtained by the nearest neighbor methods as features for an SVM learner rather than as estimates of an essay's score will answer our first objection. Our second objection is answered by the fact that our system will use the estimates given by all the nearest neighbor methods rather than only one of them. Finally,

because the regressor outputs a real-valued organization score, we round its results to the nearest of the seven organization scores the annotators were permitted to assign.

One final question we have is: Can we help our SVM regressor predict an essay's organization score by providing it with additional knowledge not derived through the use of nearest neighbor methods? Recall we mentioned that an essay's organization score is closely related to the sequence of paragraphs it contains. The best way to use paragraph sequences to predict organization scores is not clear, however. We have already described using features derived indirectly from paragraph sequences through sequence alignment and nearest neighbor approaches. Is there a way that we can make more direct use of this information?

One simple way we can extract additional features is to construct paragraph label n-grams from the predicted paragraph sequence labels of both the training and test essays. To illustrate the intuition behind these features, consider two paragraph sequence bigrams: Introduction-Body and Rebuttal-Introduction. It is fairly typical to see the first bigram, I-B, at the beginning of a good essay, so its occurrence should give us a small amount of evidence that the essay it occurs in is organized well. The presence of the second bigram, R-I, however, should indicate that its essay's organization is poor because, in general, a good essay should not give a Rebuttal before an Introduction. Because we can envision paragraph label n-grams of various lengths being useful, we create a binary presence or absence feature for each 1, 2, 3, 4, or 5-gram paragraph sequence appearing in the HMM-paragraph-labeled training set.

CHAPTER 4
MODEL EVALUATION

4.1 Evaluation Metrics

We designed three evaluation metrics to measure the error of our organization scoring system. The simplest metric, S_1, is perhaps the most intuitive. It simply measures the frequency at which a system predicts the wrong score out of the seven possible scores. Hence, a system that predicts the right score only 25% of the time would receive an S_1 score of 0.75.

The S_2 metric is slightly less intuitive than S_1, but no less reasonable. It measures the average distance between the system's score and the annotated score. This metric reflects the idea that a system that estimates organization scores close to the annotator-assigned scores should be preferred over a system whose estimations are farther off, even if both systems estimate the correct score at the same frequency.

Finally, the S_3 evaluation metric measures the average square of the distance between a system's organization score estimations and the annotator-assigned scores. The intuition behind this system is that not only should we prefer a system whose estimations are close to the annotator scores, but we should also prefer one whose estimations are not often very far away from the annotator scores. These three evaluation metrics are given by:

$$S_1 = \frac{1}{N}\sum_{A_i \neq E_i} 1, \qquad S_2 = \frac{1}{N}\sum_{i=1}^{N} |A_i - E_i|, \qquad S_3 = \frac{1}{N}\sum_{i=1}^{N} (A_i - E_i)^2$$

where A_i and E_i are the annotator assigned and system estimated scores respectively for essay i, and N is the number of essays.

4.2 Baseline Scoring Systems

Because there has been little work on this topic, there is no published model of essay organization that represents the current state of the art. Hence, both of our baseline systems are somewhat naïve. Our first baseline, B_r, assigns a test essay a score chosen uniformly at random from one of the seven possible scores. We can view B_r's performance with respect to the three scoring metrics a lower bound for any reasonable system. While it is possible to obtain scores lower than those of B_r, any scoring system performing worse than it can be viewed as actively making poor decisions.

Our second baseline B_m, is less naïve than the first in that it makes use of training data. Given a test essay, B_m assigns it the score occurring most frequently in the training set. We expect this baseline to do well largely due to the skewed distribution of essays among the seven organization scores. B_m is a useful baseline because we can view any system that performs as well or better than it as actively making good decisions regarding score assignments.

4.3 Experimental Results

We call our system R_{np} because it uses regression to predict essay scores, and it uses both nearest neighbor features and paragraph sequence features. To test our system, we performed 5-fold cross validation on our set of essays, microaveraging our results into three scores corresponding to the three scoring metrics described above. We additionally tested the significance of our system R_{np}'s results under each metric with respect to the best baseline B_m and two other tested systems R_n and R_p described later using the t-test. For our first experiment, we compared the scores obtained by our system, R_{np} to those obtained by the highest scoring baseline system, B_m. We

discovered that, while R_{np} outperforms the baseline with respect to all three metrics, the difference between the two systems was only statistically significant for metrics S_2 and S_3.

Recall that R_{np} uses features generated through nearest neighbor methods and paragraph label sequence n-grams. While it is clear that R_{np} performs better than the best performing baseline, it is important to understand how each type of feature contributed to its performance. To do this, we first remove the nearest neighbor features from our system, yielding R_p, a regression system using only paragraph sequence n-gram features. Next, we remove paragraph sequence features from our system, yielding R_n, a system using only nearest neighbor features. This method of examining each type of feature independently is motivated by feature ablation.

Table 4.1. Experimental results of organization score prediction systems

System	S_1	S_2	S_3
B_r	0.855	1.044	1.619
B_m	0.608	0.426	0.357
R_{np}	0.597	0.401$*\dagger\diamond$	0.314$*\ddagger$
R_p	0.611	0.415	0.332
R_n	0.603	0.405	0.317
H_1	0.618	0.433	0.363
H_2	0.613	0.431	0.361
H_3	0.612	0.399	0.307

$*$ significant with respect to B_m at the 0.99 level
$\dagger$ significant with respect to R_p at the 0.90 level
$\ddagger$ significant with respect to R_p at the 0.95 level
$\diamond$ significant with respect to R_n at the 0.90 level

We can see from Table 4.1 that both types of features contribute to our system's performance, as R_{np} outperforms both R_n and R_p under all three scoring metrics. The statistical significance results suggest, however, that nearest neighbor features are more important to our system's strong performance, at least under scoring

metric S_3. Under metrics S_1 and S_2, the results still show that nearest neighbor features are still more important, but not significantly so.

Next, we are interested in learning what effect the two-layer HMM has on our system. Recall that both sets of features used by our regression system rely on paragraph sequence labels. The primary aim of our two-layer HMM is to generate paragraph label sequences for both training and test essays. Since we constructed a set of heuristic rules for assigning estimated paragraph labels to the training essays, one way we can evaluate our HMM's effectiveness at this task is to run the two-layer HMM in the normal way, but rather than using the HMM-assigned paragraph labels to generate features for the regressor, use the heuristic rules on the HMM-labeled sentences, then construct the regressor's features using the heuristically labeled paragraph sequences. System H_1 does exactly that. When we compare this system to R_{np}, the results suggest that the two-layer HMM is clustering paragraphs in a more useful way for regression than do the heuristic rules.

One problem with the H_1 experiment is that it still depends on the two-layer HMM to provide the test essay sentence function labels off which the heuristically assigned paragraph labels are based. If we want to evaluate the effects of our two-layer HMM, it may be more appropriate to label the test sentences with a traditional HMM instead, then use those sentence function labels to heuristically assign paragraph labels. System H_2, which assigns paragraph and sentence function labels in this way, performs slightly worse by all evaluation metrics than B_m, providing yet more evidence that the two-layer HMM is useful for this task.

Finally, system H_3 assigns test essay paragraph labels heuristically based on the true (annotator-assigned) sentence function labels. This experiment is unrealistic in that it assumes we have access to the true function labels of test essays, but it

may still serve to give us an upper limit on how well our R_{np} system can perform if no errors were introduced by the two-layer HMM. Additionally, it may give us some insight into how well the two-layer HMM clusters paragraphs by type compared to how well the heuristic rules do it. The results show that H_3 performs comparably to R_{np} by the S_2 and S_3 evaluation metrics, but worse on the S_1 metric. Aside from reinforcing our observation that assigning test essays their exact organization scores is more difficult than assigning close estimates of the actual score, the results of this experiment suggest that it would be difficult to improve our system's performance much without either radically changing the way paragraph labels are assigned, or changing how the regressor's features are generated.

CHAPTER 5
CONCLUSION

In this thesis, we described the relatively unstudied problem of automatically scoring the organization of student essays. Since progress in modeling organization is hindered in part by the lack of a publicly available annotated corpus, we first developed a detailed essay annotation rubric that examines ten distinct dimensions of essay quality and labels the discourse function of sentences. We then annotated a corpus of student essays selected from the ICLE, which is rich in the type of discourse structures we want to model.

We then designed a computational model for scoring the organization of student essays. Our model labels discourse structures at both the local and global levels of a test essay and then relates these structures to previously seen paragraphs and essays to evaluate the test essay's organization. Experimental results show that our scoring system significantly outperforms a baseline system in two out of three metrics. Finally, we analyzed the impact of the nearest neighbor and paragraph n-gram features on our system, as well as the impact of one of our system's major components, the two-layer HMM. To stimulate further work on this problem, we make our corpus of annotated student essays available to other researchers, as we believe that our data set will be a valuable resource to the NLP community.

REFERENCES

[1] Mark D. Shermis and Jill C. Burstein. *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2003.

[2] M. D. Shermis, J. Burstein, D. Higgins, and K. Zechner. Automated essay scoring: Writing assessment and instruction. In *International encyclopedia of education (3rd edition)*. Elsevier, Oxford, UK, (in press).

[3] Y. Attali and J. Burstein. Automated essay scoring with E-rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3), 2006.

[4] Thomas K Landauer, Darrell Laham, and Peter W. Foltz. Automated scoring and annotation of essays with the Intelligent Essay Assessor^{TM}. 2003.

[5] E. Miltsakaki and K. Kukich. Evaluation of text coherence for electronic essay scoring systems. *Natural Language Engineering*, 10(1):25–55, 2004.

[6] Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. Evaluating multiple aspects of coherence in student essays. In *HLT-NAACL 2004: Main Proceedings*, pages 185–192, 2004.

[7] T. Silva. Toward an understanding of the distinct nature of L2 writing: The ESL research and its implications. 27(4):657–677, 1993.

[8] Regina Barzilay and Lillian Lee. Catching the drift: Probabilistic content models, with applications to generation and summarization. In *HLT-NAACL 2004: Main Proceedings*, pages 113–120, 2004.

[9] Regina Barzilay and Mirella Lapata. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL-05), pages 141–148, 2005.

[10] Radu Soricut and Daniel Marcu. Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 803–810, 2006.

[11] Micha Elsner, Joseph Austerweil, and Eugene Charniak. A unified local and global model for discourse coherence. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* (NAACL-HLT), pages 436–443, 2007.

[12] Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. *International Corpus of Learner English (Version 2)*. Presses universitaires de Louvain, 2009.

[13] Jill Burstein, Daniel Marcu, and Kevin Knight. Finding the write stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, 18(1):32–39, 2003.

[14] William C. Mann and Sandra A. Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[15] Nuria Oliver, Ashutosh Garg, and Eric Horvitz. Layered representations for learning and inferring office activity from multiple sensory channels. *Comput. Vis. Image Underst.*, 96(2):163–180, 2004.

[16] Yiming Yang and Jan O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of ICML*, pages 412–420, 1997.

[17] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269, January 1967.

[18] Saul Ben Needleman and Christian Dennis Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, March 1970.

[19] Thorsten Joachims. Making large-scale support vector machine learning practical, 1998.

[20] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Mach. Learn.*, 20(3):273–297, 1995.

VITA

Alan Davis was born on September 25, 1986, in Brookline, Massachusetts to parents Gabriel Davis and Roberta Leahy. He was always interested in taking things apart and putting them back together whenever he could. Legos were never enough for him; he needed to disassemble bicycles and lawnmowers to investigate how they worked. Alan spent his childhood in rural New Hampshire, where he gained an appreciation for wandering in the woods, riding all-terrain vehicles, and generally being outside. His strongest subject in school was always Mathematics, where he excelled from an early age. In high school, Alan took a Computer Science course, where he learned the Java programming language. He spent some of his free time the following summer programming fractals, the SET card game, and a 3D graphics engine from scratch. In 2004, he decided to attend the University of Texas at Dallas to study Computer Science. There he took three courses from Dr. Vincent Ng — Discrete Mathematics, Artificial Intelligence, and Machine Learning. During these courses, Alan came to know Dr. Ng through extensive conversations in his office that often lasted late into the night. After being awarded the Erik Jonsson Distinguished Scholar Fellowship, he pursued a Master's degree in Computer Science, following the Intelligent Systems track with Dr. Ng as his research advisor. In April 2010, he completed his thesis. After living in Texas for twelve years, Alan will move west to pursue his passions in Computer Science, Mathematics, and teaching.

Permanent address: 2507 Bluebonnet Road
Austin, Texas 78704